the URLs. Administrative tools 120 permit the operator to configure various operating parameters.

The pattern recognizer 116 and scheduler 118 cooperate to enable intelligent pre-caching of frequently requested content. The operation of the local service provider 110 to perform this intelligent pre-caching according to an aspect of this invention is described in conjunction with reference to the flow diagram of Fig. 5. The local service provider is programmed to perform the computer-implemented steps of Fig. 5 to alleviate the problems of providing streaming video and audio data over the Internet. The steps are presented in the illustrated order for discussion purposes, but are not restricted to this sequence.

The pattern recognizer 116 monitors the patterns of the subscriber requests to determine which content is most frequently requested and when (step 150 in Fig. 5). From these patterns, the pattern recognizer 116 can identify peak times in subscriber traffic and the relation of the peak times to specific requested content (step 152). For instance, suppose that a high number of subscribers frequently request the CNN Web page during the morning hours of 6:30 AM to 8:00 AM. These requests translate into a high number of URL hits for the CNN Web page which are recorded by hit recorder 112 in the URL hit database 114. The pattern recognizer 116 recognizes this recurring pattern of requests for the CNN Web page and identifies the peak time for this Web page to be between 6:30 AM and 8:00 AM.

Using the patterns identified by the pattern recognizer 116, the scheduler 118 schedules delivery of the content at a selected time prior to the peak time (step 154 in Fig. 5). In this example, the scheduler 118 might schedule delivery of the CNN Web page at a time prior to 6:30 AM. For instance, the scheduler 118 might schedule a request for the CNN Web page at 6:00 AM to provide sufficient time to downloaded that page before the earliest subscribers are expected to begin asking for it, yet not too early to ensure that the latest news is included.

At the scheduled time, a media loader 122 sends a request to the content server on the Internet and receives the content from that content server (step 156 in Fig. 5). The content is stored locally at the local service provider (step 158). More particularly, the data comprising the target resource is stored as a proxy file in the cache memory 124, and any continuous data content (e.g., audio or video data) is stored in the continuous media server 126. In the Web context, the content might be in the form of a Web page or other hypermedia document that has hyperlinks to various data items, such as audio and/or video clips. The hypermedia document itself is stored in the cache memory 124, while the audio and video clips referenced by the hyperlinks are stored in the CMS 126. The target specifications corresponding to the links in the cached hypermedia document are modified to reference the audio and video files in the CMS 126, as opposed to the files maintained at the Web site (step 160 in Fig. 5). As an alternative to modifying the target specifications, a conversion table can be constructed which converts requests from referencing the files at the Web site to referencing the files in the CMS 126.

In our CNN example, the local service provider 110 sends a request to the CNN Web site seeking to download the CNN Web page at 6:00 AM. The CNN Web page is downloaded over the Internet and stored in the cache memory 124. If the CNN Web page contains links to any audio or video clips of recent news, these data files are also downloaded and stored in the CMS 126. The links within the cached Web page are modified to reference the audio and video files stored locally in the CMS 126, instead of the files maintained at the CNN Web site.

The media loader 122 loads the locally stored content just before the peak time so that it is ready to serve during the peak time (step 162 in Fig. 5). When the first subscriber requests the CNN Web page at, for example, 6:40 AM, the local service provider 110 serves the Web page from the cached memory 124. If the subscriber activates a link to a video or audio file, the local service provider 110 immediately serves the data stream from the CMS 126 for just-in-time rendering on the subscriber's computer. Accordingly, the video file is served as streaming data to even the first subscriber who requests it, rather than making that subscriber wait for the file to be retrieved over the Internet.

The intelligent pre-caching method obviates the latency problems associated with streaming video and audio over the Internet, and is a further improvement to traditional on-demand caching techniques. However, it is noted that the network system does not accommodate data streaming for every video and audio file on the Internet, but instead only selected files. The system makes an intelligent choice as to which content is likely to be requested by its subscribers and then makes only this content readily available to the subscribers. In this way, the method seeks to optimize the physical computing resources of the local service provider in a manner which best services the majority of the clientele.

It is noted that the content servers serve many local service providers over the Internet. These local service providers, in turn, serve many different users. Due to varying demographics, the local service providers will generally differ in the content that it most often serves to its clientele. For example, a service provider in Seattle, Washington, might have many requests for content on entertainment or news local to Seattle. The pattern recognizer for a Seattle-based service provider might therefore schedule proportionally more Seattle related content than, say, a London-based service provider. As a result, the sets of pre-cached content may differ significantly from one service provider to another depending upon the results of the local hit recorder and pattern recognizer.

The system and method described above places the authority for deciding which content is pre-cached at the local service provider. This allows the local service provider to adapt to the often changing patterns of its clientele. However, in another implementation, the content servers can be given the governing authority of deciding when and what content to download to the ISPs prior to peak times. For instance, the content server can maintain a schedule of when to download different sets of content to various ISPs in timely fashion before the sets of content are requested by the respective users who are serviced by the ISPs.

With continuing reference to Fig. 4, in this implementation, the local service provider 110 also includes a policy manager 128 which defines and administers rules that determine which documents or resources are cached in the cache memory 124. For instance, caching rules might call for caching resources that are routinely requested by many subscribers, but foregoing caching resources that are rarely or infrequently requested. The policy rules also coordinate cache maintenance by deciding when documents are out-of-date and how these documents are deleted from the cache memory 124.

According to another aspect of this invention, time-to-live (TTL) tags are assigned to the content to assist in determining when the content should be refreshed or disposed. The TTL tags can be assigned by the content server as part of the content itself. The server can attach an expiration tag which represents the publisher's best estimate as to how long it will be before the content is updated.

Alternatively, the local service providers might compute the TTL tags for the content it caches in cache memory 124. The computation is based upon a theory that older content is less likely to change. Content that changed only 10 minutes ago is statistically more likely to change within the next 24 hours than content that last changed one month ago. In one implementation, an approximate TTL is computed as a percentage of time since the content is known to have last changed. The percentage is an operator controlled parameter. Suppose a 10% value is selected. Content that last changed 72 hours ago is assigned a TTL tag of 7.2 hours. If the content is not updated within 7.2 hours, it is given a new value of 7.9 hours (i.e., (72 hours + 7.2 hours) x 10% = 7.9 hours). As the content ages, it is checked less often. The TTL tags can be kept in a separate table of the cache 124 to correlate the tags and their content.

Deletion policies are a function of the content itself (e.g., its TTL tags), the subscriber patterns (e.g., how frequently the content is requested), the cost to request newer updated content, and the constraints imposed by capacity limitations of the cache memory.

The local service provider 110 also maintains a subscriber database 130 which stores lists of subscribers (or LAN users in the LAN configuration) and pertinent information about them (e.g., routing addresses, billing addresses, etc.). A usage reporter 132 uses the URL hit information from the URL hit database 114 and subscriber information from the subscriber database 130 to generate reports on subscriber usage patterns. These reports can be used by the operator to efficiently allocate computer resources to best satisfy the needs of its clientele. The reports can also be used by content providers to help them assess the popularity of their Web sites and the type of subscribers who visit them.

In a preferred implementation, the functional components described with respect to Fig. 4 are implemented in software which executes on the host computer of the local service provider. It is noted that the functional layout is provided for explanation purposes. The subscriber database 130, the URL hit database 114, and the cache memory 124 can be implemented as one database server. Other implementation variations may also be made.

In the above system, the local service providers (e.g., ISPs, LAN Web servers) initiate the requests for content so that it may be pre-cached prior to peak demand times when the content is most likely to be requested. This system can be referred to as a "pull-caching" system in that content is pulled over the Internet upon request of the local service providers. The method of intelligently pull-caching data prior to peak times enables delivery of streaming video and audio data to Internet users.

Fig. 6 shows a network system 200 according to another aspect of this invention. The network system 200 attacks the latency problem of streaming video and audio data by supplementing the primary Internet distribution network with a second network which is not reliant on the Internet/ISP connection. The Internet/ISP connection is often the bottleneck for streaming data and is typically the connection least likely to be upgraded due to economic factors surrounding the business of the ISP. Although not required, in this implementation, the content may be pushed top down from the content provider over the Internet and thus, the system may be referred to as a "push-caching" system.

Network system 200 is similar to the configuration of the Fig. 2 network system 50 in that it has a content server 52 which serves content over a high-speed, high-bandwidth network 54, via local ISPs 56, to end users 58 and 60. The difference between the two systems is that network system 200 of Fig. 6 has an additional, secondary network 202 for distributing content from the content server 52 to the ISPs 56. In the illustrated implementation, the secondary network 202 is a broadcast satellite network. The content provider 52 has a transmitter 204 which sends signals to an orbiting satellite 206, which redirects the signals to an ISP-based receiver 208.

The secondary satellite network 202 affords a supplemental bandwidth for delivery of content to participating ISPs in addition to the content delivered over the interactive network connection 62. For instance, using present DSS (digital satellite service) technology, the satellite network 202 provides an additional 6 Mbps bandwidth to deliver content to the ISP 56. This extra bandwidth is made available at a fraction of the cost of buying T1/T3 connections.

The supplemental-caching technique allows the content provider to download more information in a timely manner. To continue the above CNN example, the CNN content provider might transmit the CNN Web page over the satellite system 202 minutes before the peak time of 6:30 AM to 8:00 AM. The ISP receives the Web page from its satellite receiver 208 and caches the Web page for serving during the peak time. The CNN Web page is thereby efficiently made available for real-time streaming to the subscribers, without tying up or consuming any of the bandwidth provided by the network connection 62.

In a preferred implementation, the supplemental satellite network 202 is uni-directional, in that data is broadcast from the content provider 52 to the ISP

56. It is a low cost solution increasing the bandwidth of the pipeline to the ISP, without requiring significant investment on the part of the ISP. In addition to satellite technologies, the broadcast network 202 can be implemented as other wireless systems, such as RF or cellular technologies.

In another embodiment, the secondary network 202 can be implemented as an second data communications network whereby supplemental content is multicasted to participating ISPs prior to the peak time.

In compliance with the patent statutes, the invention has been described in language more or less specific as to structure and method features. It is to be understood, however, that the invention is not limited to the specific features described, since the means herein disclosed comprise exemplary forms of putting the invention into effect. The invention is, therefore, claimed in any of its forms or modifications within the proper scope of the appended claims appropriately interpreted in accordance with the doctrine of equivalents and other applicable judicial doctrines.